Sloan School of Management
Massachusetts Institute of Technology
Cambridge 39, Massachusetts
December, 1964

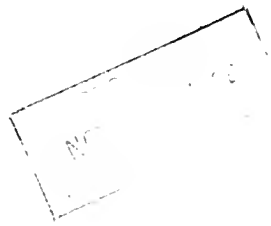
Multicollinearity in Regression Analysis: The Problem Revisited

105-64

D. E. Farrar and R. R. Glauber*

 The research has been supported by the Institute of Naval Studies, and by grants from the Ford Foundation to the Sloan School of Management, and the Harvard Business School.

*Sloan School of Management, M.I.T., and Graduate School of Business Administration, Harvard University, respectively.



CONTENTS

MULTICOLLINEARITY IN REGRESSION ANALYSIS: THE PROBLEM REVISITED

To most economists the single equation least squares regression model, like an old friend, is tried and true. Its properties and limitations have been extensively studied, documented and are, for the most part, well known. Any good text in econometrics can lay out the assumptions on which the model is based and provide a reasonably coherent -- perhaps even a lucid -- discussion of problems that arise as particular assumptions are violated. A short bibliography of definitive papers on such classical problems as non-normality, heteroscedasticity, serial correlation, feedback, etc., completes the job.

As with most old friends, however, the longer one knows least squares, the more one learns about it. An admiration for its robustness under departures from many assumptions is sure to grow. The admiration must be tempered, however, by an appreciation of the model's sensitivity to certain other conditions. The requirement that independent variables be truly independent of one another is one of these.

Proper treatment of the model's classical problems ordinarily involves two separate stages, detection and correction. The Durbin-Watson test for serial correlation, combined with Cochrane and Orcutt's suggested first differencing procedure, is an obvious example.*

---

*J. Durbin and G. S. Watson, "Testing for Serial Correlation in Least Squares Regression, Biometrika, 37-8, 1950-1. D. Cochrane and G. H. Orcutt. "Application of Least Squares Regressions to Relationships Containing Auto-Correlated Error Terms," J. Am. Statistical Assoc., 44, 1949.

Bartlett's test for variance heterogeneity followed by a data transformation to restore homoscedasticity is another.* No such "proper treatment" has been developed, however, for problems that arise as multicollinearity is encountered in regression analysis.

Our attention here will focus on what we consider to be the first step in a proper treatment of the problem -- its detection, or diagnosis. Economists generally agree that the second step -- correction -- requires the generation of additional information.** Just how this information is to be obtained depends largely on the tastes of an investigator and on the specifics of a particular problem. It may involve additional primary data collection, the use of extraneous parameter estimates from secondary data sources, or the application of subjective information through constrained regression, or through Bayesian estimation procedures. Whatever its source, however, selectivity -- and thereby efficiency -- in generating the added information requires a systematic procedure for detecting its need -- i.e., for detecting the existence, measuring the extent, and pinpointing the location and causes of multicollinearity within a set of independent variables. Measures are proposed here that, in our opinion, fill this need.

The paper's basic organization can be outlined briefly as follows. In the next section the multicollinearity problem's basic, formal nature is developed and illustrated. A discussion

---

*F. David and J. Neyman, "Extension of the Markoff Theorem on Least Squares," Statistical Research Memoirs, II. London, 1938.

**J. Johnston, Econometric Methods, McGraw Hill, 1963, p. 207; J. Meyer and R. Glauber, Investment Decisions, Economic Forecasting, and Public Policy, Division of Research Graduate School of Business Administration, Harvard University, 1964, p. 181 ff.

of historical approaches to the problem follows. With this as background, an attempt is made to define multicollinearity in terms of departures from a hypothesized statistical condition, and to fashion a series of hierarchical measures -- at each of three levels of detail -- for its presence, severity, and location within a set of data. Tests are developed in terms of a generalized, multivariate normal, linear model, A pragmatic interpretation of resulting statistics as dimensionless measures of correspondence between hypothesized and sample properties, rather than in terms of classical probability levels, is advocated. A numerical example and a summary completes the exposition.

THE MULTICOLLINEARITY PROBLEM

NATURE AND EFFECTS

The purpose of regression analysis is to estimate the parameters of a dependency, not of an interdependency, relationship. We define first

- $\underline{Y}$, $\underline{X}$ as observed values, measured as standardized deviates, of the dependent and independent variables,
- $\underline{\beta}$ as the true (structural) coefficients,
- $\underline{u}$ as the true (unobserved) error term, with distributional properties specified by the general linear model,* and
- σ_u^2 as the underlying, population variance of $\underline{u}$;

and presume that $\underline{Y}$ and $\underline{X}$ are related to one another through the linear form

$$\underline{Y} = \underline{X}\,\beta + \underline{u} \quad . \tag{1}$$

Least squares regression analysis leads to estimates

$$\hat{\underline{\beta}} = (\underline{X}^t\underline{X})^{-1}\,\underline{X}^t\,\underline{Y}$$

with variance-covariance matrix

$$\text{var}(\hat{\underline{\beta}}) = \sigma_u^2\,(\underline{X}^t\underline{X})^{-1} \quad ,$$

---

*See for example, J. Johnston, op.cit., Ch. 4; or F. Graybill, An Introduction to Linear Statistical Models, McGraw Hill, 1961, Ch. 5.

that, in a variety of senses, best reproduces the the hypothesized dependency relationship (1).

Multicollinearity, on the other hand, is veiwed here as an interdependency condition. It is defined in terms of a lack of independence, or of the presence of interdependence -- signified by high intercorrelations ($\underline{X}^t\underline{X}$) -- within a set of variables, and under this view can exist quite apart from the nature, or even the existence of a dependency relationship between $\underline{X}$ and a dependent variable $\underline{Y}$. Multicollinearity is not important to the statistician for its own sake. Its significance, as contrasted with its definition, comes from the effect of interdependence in $\underline{X}$ on the dependency relationship whose parameters are desired. Multicollinearity constitutes a threat -- and often a very serious threat -- both to the proper specification and to the effective estimation of the type of structural relationships commonly sought through the use of regression techniques.

The single equation, least squares regression model is not well equipped to cope with interdependent explanatory variables. In its original and most simple form the problem is not even conceived. Values of $\underline{X}$ are presumed to be the pre-selected, controlled elements of a classical, laboratory experiment.* Least squares models are not limited, however, to simple, fixed variate -- or fully controlled -- experimental situations. Partially controlled or completely uncontrolled experiments, in which $\underline{X}$ as well as $\underline{Y}$ is subject to random variation -- and therefore also to multicollinearity -- may provide the data on which perfectly legitimate regression analyses are based.**

---

*Kendall puts his finger on the essence of the simple, fixed variate, regression model when he remarks that standard multiple regression is not multivariate at all, but is univariate. Only $\underline{Y}$ is presumed to be generated by a process that includes stochastic elements. M. G. Kendall, A Course in Multivariate Analysis, Hafner, 1957, p. 68-9.

**See Model 3, Graybill, op. cit. p. 104.

Though not limited to fully controlled, fixed variate experiments, the regression model, like any other analytical tool, is limited to good experiments if good results are to be insured. And a good experiment must provide as many dimensions of independent variation in the data it generates as there are in the hypothesis it handles. An n dimensional hypothesis -- implied by a regression equation containing n independent variables -- can be neither properly estimated nor properly tested, in its entirety, by data that contains fewer than n significant dimensions.

In most cases an analyst may not be equally concerned about the structural integrity of each parameter in an equation. Indeed, it is often suggested that for certain forecasting applications structural integrity anywhere -- although always nice to have -- may be of secondary importance. This point will be considered later. For the moment it must suffice to emphasize that multicollinearity is both a symptom and a facet of poor experimental design. In a laboratory poor design occurs only through the improper use of control. In life, where most economic experiments take place, control is minimal at best, and multicollinearity is an ever present and seriously disturbing fact of life.

Estimation

Difficulties associated with a multicollinear set of data depend, of course, on the severity with which the problem is encountered. As interdependence among explanatory variables $\underline{X}$ grows, the correlation matrix $(\underline{X}^t\underline{X})$ approaches singularity, and elements of the inverse matrix $(\underline{X}^t\underline{X})^{-1}$ explode. In the limit, perfect linear dependence within an independent variable set leads to perfect singularity on the part of $(\underline{X}^t\underline{X})$ and to a completely indeterminate set of parameter estimates $\hat{\underline{\beta}}$. In a formal sense diagonal elements of the inverse matrix $(\underline{X}^t\underline{X})^{-1}$ that correspond to linearly dependent members of $\underline{X}$ become infinite. Variances for the affected variables' regression coefficients,

$$\text{var}(\underline{\hat{\beta}}) = \sigma_u^2 \, (\underline{X}^t \underline{X})^{-1} \quad ,$$

accordingly, also become infinite.

The mathematics, in its brute and tactless way, tells us that explained variance can be allocated completely arbitrarily between linearly dependent members of a completely singular set of variables, and almost arbitrarily between members of an almost singular set. Alternatively, the large variances on regression coefficients produced by multicollinear independent variables indicate, quite properly, the low information content of observed data and, accordingly, the low quality of resulting parameter estimates. It emphasizes one's inability to distinguish the independent contribution to explained variance of an explanatory variable that exhibits little or no truly independent variation.

In many ways a person whose independent variable set is completely interdependent may be more fortunate than one whose data is almost so; for the former's inability to base his model on data that cannot support its information requirements will be discovered -- by a purely mechanical inability to invert the singular matrix $(\underline{X}^t \underline{X})$ -- while the latter's problem, in most cases, will never be fully realized.

Difficulties encountered in the application of regression techniques to highly multicollinear independent variables can be discussed at great length, and in many ways. One can state that the parameter estimates obtained are highly sensitive to changes in model specification, to changes in sample coverage, or even to changes in the direction of minimization; but lacking a simple illustration, it is difficult to endow such statements with much meaning.

Illustrations, unfortunately, are plentiful in economic research. A simple Cobb-Douglas production function provides an almost classical example of the instability of least squares parameter estimates when derived from collinear data.

Illustration

The Cobb-Douglas production function* can be expressed most simply as

$$P = L^{\beta 1} C^{\beta 2} e^{\alpha+u}$$

where

P is production or output,

L is labor input,

C is captial input,

$\alpha, \beta_1 \beta_2$ are parameters to be estimated, and

u is an error or residual term.

Should $\beta_1+\beta_2 = 1$, proportionate changes in inputs generate equal, proportionate changes in expected output -- the production function is linear, homogeneous -- and several desirable conditions of welfare economics are satisfied. Cobb and Douglas set out to test the homogeneity hypothesis empirically. Structural estimates, accordingly, are desired.

Twenty-four annual observations on aggregate employment, capital stock and output for the manufacturing sector of the U.S. economy, 1899 and 1922, are collected. β_2 is set equal to $1-\beta_1$ and the value of the labor coefficient $\beta_1 = .75$ is estimated by constrained least squares regression analysis. By virtue of the constraint, $\beta_2 = 1 - .75 = .25$. Cobb and Douglas are satisfied that structural estimates of labor and capital coefficients for the manufacturing sector of the U.S. economy have been obtained.

---

* C. W. Cobb and P. H. Douglas, "A Theory of Production," American Economic Review, XVIII, Supplement, March 1928. See H. Mendushausen, "On the Significance of Professor Douglas' Production Function," Econometrica, 6 April 1938; and D. Durand, "Some Thought on Marginal Productivity with Special Reference to Professor Douglas Analysis," Journal of Political Economy, 45, 1937.

Ten years later Mendershausen reproduces Cobb and Douglas' work and demonstrates, quite vividly, both the collinearity of their data and the sensitivity of results to sample coverage and direction of minimization. Using the same data we have attempted to reconstruct both Cobb and Douglas' original calculations and Mendershausen's replication of same. A demonstration of least squares' sensitivity to model specification -- i.e., to the composition of an equation's independent variable set -- is added. By being unable to reproduce several of Mendershausen's results, we inadvertently (and facetiously) add "sensitivity to computation error" to the table of pitfalls commonly associated with multicollinearity in regression analysis. Our own computations are summarized below.

Parameter estimates for the Cobb-Douglas model are linear in logarithms of the variables P, L, and C. Table 1 contains simple correlations between these variables, in addition to an arithmetic trend, t. Multicollinearity within the data is indicated by the high intercorrelations between all variables, a common occurrence when aggregative time series data are used.

The sensitivity of parameter estimates to virtually any change in the delicate balance obtained by a particular sample of observations, and a particular model specification, may be illustrated forcefully by the wildly fluctuating array of labor and capital coefficients, β_1 and β_2, summarized in Table 2.

Equations (a) and (b) replicate Cobb and Douglas' original, constrained estimates of labor and capital coefficients with and without a trend to pick up the impact on productivity of technological change.* Cobb and Douglas comment on the in-

---

* The capital coefficient β_2 is constrained to equal $1-\beta_1$ by regressing the logarithm of P/C on the logarithm of L/C, with and without a term for trend, $\beta_3 t$.

crease in correlation that results from adding a trend to their production function,* but neglect to report the change in production coefficients that accompany the new specification (equation b, Table 2).

Table 1
Simple Correlations,
Cobb-Douglas Production Function

	Log P	Log L	Log C	t
Log P	1.00			
Log L	.96	1.00		
Log C	.95	.91	1.00	
t	.94	.90	.99*	1.00

* To four places, $r_{C,t} = .9968$

The sensitivity of coefficient estimates to changes in model specification is demonstrated even more strikingly, however, by equations (c) and (d), Table 2. Here estimates of the logarithm of production are based directly on logarithms of labor and capital, with and without a term for trend. β_1 and β_2, accordingly, are not constrained to add to unity. The relationship obtained in equation (c) is very nearly linear, homogeneous -- $\beta_1 + \beta_2 = 1.04$. Cobb and Douglas, clearly, would be delighted by this specification. Unconstrained parameters closely resemble their own constrained estimates of labor and

* Cobb, Douglas, op cit, p. 154.

Table 2

Regression Coefficients*
Cobb-Douglas Production Relationship

Equation Number	Dependent Variable, Estimation Equation	β_1 Labor	β_2 Capital	β_3 Trend	R^2
	Constrained Estimation				
a.	Log P/C	.75(18.1)	.25	--	.93
b.	Log P/C	.89(6.0)	.11	.01(1.0)	.93
	Unconstrained Least Squares				
c.	Log P	.81(5.6)	.23(3.7)	--	.95
d.	Log P	.91(6.5)	-.53(1.5)	.05(2.3)	.96
e.	Log C	.05	.60	--	.88
f.	Log L	1.35	.01	--	.92
	Variations in Sample Coverage Yrs Excluded				
g. 1908	Log P	.75(5.0)	.25(4.0)	--	.95
h. 1921	Log P	.60(3.3)	.34(4.0)	--	.96
i. 1922	Log P	1.02(8.0)	.12(2.0)	--	.97

* $t = \beta/\hat{\sigma}_\beta$ follow regression coefficients in parentheses -- not calculated for the indirectly estimated parameters in equations a, b, e, f.

R^2, t-ratios are corrected for degrees of freedom.

capital coefficients, and the observed relationship between dependent and independent variables is strong, both individually (as indicated by t-ratios), and jointly (as measured by R^2, the coefficient of determination).

By adding a trend to the independent variable set, however the whole system rotates wildly. The capital coefficient $\beta_2 = .23$ that makes equation (c) so reassuring becomes $\beta_2 = -.53$ in equation (d). Labor and technological change, of course, pick up and divide between themselves much of capital's former explanatory power. Neither, however, assumes a value that, on a priori grounds, can be dismissed as outrageous; and, it must be noted, each variable's individual contribution to explained variance (measured by t-ratios) continues to be strong. Despite the fact that both trend and capital -- and labor too, for that matter -- carry large standard errors, no "danger light" is flashed by conventional statistical measures of goodness of fit. Indeed, R^2 and t-ratios for individual coefficients are sufficiently large in either equation to lull a great many econometricians into a wholly unwarranted sense of security.

Evidence of the instability of multicollinear regression estimates under changes in the direction of minimization is illustrated by equations (c) - (f), Table 2. Here capital and labor, respectively, serve as dependent variables for estimation purposes, after which the desired (labor and capital) coefficients are derived algebraically. Labor coefficients, $\beta_1 = .05$ and 1.35, and capital coefficients, $\beta_2 = .60$ and .01, are derived from the same data that produces Cobb and Douglas' $\beta_1 = .75$, $\beta_2 = .25$ division of product between labor and capital.

Following Mendershausen's lead equations (g) - (i), Table 2, illustrate the model's sensitivity to the few non-collinear observations in the sample. By omitting a single year (1908,

21, 22 in turn) labor and capital coefficients carrying substantially different weights are obtained.

The lesson to be drawn from this exercise, of course, is that stable parameter estimates and meaningful tests of two dimensional hypotheses cannot be based on data that contains only one independent dimension. Ironically, by constraining their own coefficients to add to one -- and thereby reducing their independent variable set to a single element -- Cobb and Douglas themselves provide the only equation in Table 2 whose information requirements do not exceed the sample's information content. Equation (a) demands, and the sample provides, one and only one dimension of independent variation.*

Specification

Although less dramatic and less easily detected than instability in parameter estimates, problems surrounding model specification with multicollinear data are not less real. Far from it, correct specification is ordinarily more important to successful model building than the selection of a "correct" estimating procedure.

Poor forecasts of early postwar consumption expenditures are an example. These forecasts could have been improved marginally, perhaps, by more fortunate choices of sample, computing algorithm, direction of minimization, functional form, etc. But their basic shortcoming consists of a failure to recognize the importance of liquid assets to consumer behavior. No matter how cleverly a consumption function's coefficients are estimated, if it does not include liquid assets it cannot provide a satis-

* Lest we should be too easy on Cobb and Douglas it must be reiterated that their model embodies a two dimensional -- not a one dimensional -- hypothesis.

factory representation of postwar United States consumer behavior. Multicollinearity, unfortunately, contributes to difficulty in the specification as well as in the estimation of economic relationships.

Model specification ordinarily begins in the model builder's mind. From a combination of theory, prior information, and just plain hunch, variables are chosen to explain the behavior of a given dependent variable. The job, however, does not end with the first tentative specification. Before an equation is judged acceptable it must be tested on a body of empirical data. Should it be deficient in any of several respects, the specification -- and thereby the model builder's "prior hypothesis" -- is modified and tried again. The process may go on for some time. Eventually discrepancies between prior and sample information are reduced to tolerable levels and an equation acceptable in both respects is produced.

In concept the process is sound. In practice, however, the econometrician's mind is more fertile than his data, and the process of modifying a hypotheses consists largely of paring-down rather than of building-up model complexity. Having little confidence in the validity of his prior information, the economist tends to yield too easily to a temptation to reduce his model's scope to that of his data.

Each sample, of course, covers only a limited range of experience. A relatively small number of forces are likely to be operative over, or during, the subset of reality on which a particular set of observations is based. As the number of variables extracted from the sample increases, each tends to measure different nuances of the same, few, basic factors that are present. The sample's basic information is simply spread more and more thinly over a larger and larger number of in-

creasingly multicollinear independent variables.

However real the dependency relationship between $\underline{Y}$ and each member of a relatively large independent variable set $\underline{\underline{X}}$ may be, the growth of interdependence within $\underline{\underline{X}}$ as its size increases rapidly decreases the stability -- and therefore the sample significance -- of each independent variable's contribution to explained variance. As Liu points out, data limitations rather than theoretical limitations are largely responsible for a persistent tendency to underspecify -- or to oversimplify -- econometric models.* The increase in sample standard errors for multicollinear regression coefficients virtually assures a tendency for relevant variables to be discarded incorrectly from regression equations.

The econometrician, then, is in a box. Whether his goal is to estimate complex structural relationships in order to distinguish between alternative hypotheses, or to develop reliable forecasts, the number of variables required is likely to be large, and past experience demonstrates with depressing regularity that large numbers of economic variables from a single sample space are almost certain to be highly intercorrelated. Regardless of the particular application, then, the essence of the multicollinearity problem is the same. There exists a substantial difference between the amount of information required for the satisfactory estimation of a model and the information contained in the data at hand.

If the model is to be retained in all its complexity, solution of the multicollinearity problem requires an augmentation of existing data to include additional information. Parameter estimates for an n dimensional model -- e.g., a

---

* T. C. Liu, "Underidentification, Structural Estimation and Forecasting," Econometrica, 28, October 1960; p. 856.

two dimensional production function -- cannot properly be based on data that contains fewer significant dimensions. Neither can such data provide a basis for discriminating between alternative formulations of the model. Even for forecasting purposes the econometrician whose data is multicollinear is in an extremely exposed position. Successful forecasts with multicollinear variables require not only the perpetuation of a stable dependency relationship between $\underline{Y}$ and $\underline{\underline{X}}$, but also the perpetuation of stable interdependency relationships within $\underline{\underline{X}}$. The second condition, unfortunately, is met only in a context in which the forecasting problem is all but trivial.

The alternative of scaling down each model to fit the dimensionality of a given set of data appears equally unpromising. A set of substantially orthogonal independent variables can in general be specified only by discarding much of the prior theoretical information that a researcher brings to his problem. Time series analyses containing more than one or two independent variables would virtually disappear, and forecasting models too simple to provide reliable forecasts would become the order of the day. Consumption functions that include either income or liquid assets -- but not both -- provide an appropriate warning.

There is, perhaps a middle ground. All the variables in a model are seldom of equal interest. Theoretical questions ordinarily focus on a relatively small portion of the independent variable set. Cobb and Douglas, for example, are interested only in the magnitude of labor and capital coefficients, not in the impact on output of technological change. Disputes concerning alternative consumption, investment, and cost of capital models similarly focus on the relevance of, at most, one or two disputed variables. Similarly, forecasting models

rely for success mainly on the structural integrity of those variables whose behavior is expected to change. In each case certain variables are strategically important to a particular application while others are not.

Multicollinearity, then, constitutes a problem only if it undermines that portion of the independent variable set that is crucial to the analysis in question -- labor and capital for Cobb and Douglas; income and liquid assets for postwar consumption forecasts. Should these variables be multicollinear, corrective action is necessary. New information must be obtained. Perhaps it can be extracted by stratifying or otherwise reworking existing data. Perhaps entirely new data is required. Wherever information is to be sought, however, insight into the pattern of interdependence that undermines present data is necessary if the new information is not to be similarly affected.

Current procedures and summary statistics do not provide effective indications of multicollinearity's presence in a set of data, let alone the insight into its location, pattern, and severity that is required if a remedy -- in the form of selective additions to information -- is to be obtained. The current paper attempts to provide appropriate "diagnostics" for this purpose.

Historical approaches to the problem will both facilitate exposition and complete the necessary background for the present approach to multicollinearity in regression analysis.

HISTORICAL APPROACHES

Historical approaches to multicollinearity may be organized in any of a number of ways. A very convenient organization reflects the tastes and backgrounds of two types of persons who have worked actively in the area. Econometricians tend to view the problem in a relatively abstract manner. Computer programmers, on the other hand, see multicollinearity as just one of a relatively large number of contingencies that must be anticipated and treated. Theoretical statisticians, drawing their training, experience and data from the controlled world of the laboratory experiment, are noticeably uninterested in the problem altogether.

Econometric

Econometricians typically view multicollinearity in a very matter-of-fact -- if slightly schizophrenic -- fashion. They point out on the one hand that least squares coefficient estimates,

$$\hat{\underline{\beta}} = \underline{\beta} + (\underline{\underline{X}}^t\underline{\underline{X}})^{-1}\,\underline{\underline{X}}^t\underline{u} \quad ,$$

are "best linear unbiased," since the expectation of the last term is zero regardless of the degree of multicollinearity inherent in $\underline{\underline{X}}$, if the model is properly specified and feedback is absent. Rigorously demonstrated, this proposition is often a source of great comfort to the embattled practitioner. At times it may justify compacency.

On the other hand, we have seen that multicollinearity imparts a substantial bias toward incorrect model specification.*

---

* Liu, idem.

It has also been shown that poor specification undermines the "best linear unbiased" character of parameter estimates over multicollinear, independent variable sets.* Complacency, then, tends to be short lived, giving way alternatively to despair as the econometrician recognizes that non-experimental data, in general, is multicollinear and that "... in principal nothing can be done about it."** Or, to use Jack Johnston's words, one is "... in the statistical position of not being able to make bricks without straw."*** Data that does not possess the information required by an equation cannot be expected to yield it. Admonitions that new data or additional a priori information are required to "break the multicollinearity deadlock"**** are hardly reassuring, for the gap between information on hand and information required is so often immense.

Together the combination of complacency and despair that characterizes traditional views tends to virtually paralyze efforts to deal with multicollinearity as a legitimate and difficult, yet tractable econometric problem. There are, of course, exceptions. Two are discussed below.

Artificial Orthogonalization: The first is proposed by Kendall,***** and illustrated with data from a demand study by Stone.****** Employed correctly, the method is an example of a solution to the multicollinearity problem that proceeds by reducing a model's information requirements to the information

---

* H. Theil, "Specification Errors and the Estimation of Economic Relationships," Review of the International Statistical Statistical Institute, 25, 1957.

** H. Theil, Economic Forecasts and Policy, North Holland, 1962; p. 216.

*** J. Johnston, op. cit., p. 207.

**** J. Johnston, idem., and H. Theil, op. cit., p. 217.

***** M. G. Kendall, op. cit., pp. 70-75.

******J. R. N. Stone, "The Analysis of Market Demand," Journal of the Royal Statistical Society, CVIII, III, 1945; pp. 286-382.

content of existing data. On the other hand, perverse applications lead to parameter estimates that are even less satisfactory than those based on the original set of data.

Given a set of interdependent explanatory variables $\underline{X}$, and a hypothesized dependency relationship

$$\underline{Y} = \underline{\underline{X}}\ \underline{\beta} + \underline{u} \ ,$$

Kendall proposes "[to throw] new light on certain old but unsolved problems; particularly (a) how many variables do we take? (b) how do we discard the unimportant ones? and (c) how do we get rid of multicollinearities in them?"*

His solution, briefly, runs as follows: Defining

$\underline{\underline{X}}$ as a matrix of observations on n multicollinear explanatory variables,

$\underline{\underline{F}}$ as a set of $m \leq n$ orthogonal components or common factors,

$\underline{\underline{U}}$ as a matrix of n derived residual (or unique) components, and

$\underline{\underline{A}}$ as the constructed set of $(m \times n)$ normalized factor loadings,

Kendall decomposes $\underline{\underline{X}}$ into a set of statistically significant orthogonal common factors and residual components $\underline{\underline{U}}$ such that

$$\underline{\underline{X}} = \underline{\underline{F}}\ \underline{\underline{A}} + \underline{\underline{U}}$$

exhausts the sample's observed variation. $\hat{\underline{\underline{X}}} = \underline{\underline{F}}\underline{\underline{A}}$, then, sum-

---

* Kendall, op. cit., p. 70.

marizes $\underline{\underline{X}}$'s common or "significant" dimensions of variation in $m \leq n$ artificial, orthogonal variates, while $\underline{\underline{U}}$ picks up what little -- and presumably unimportant -- residual variation remains.

Replacing the multicollinear set $\underline{\underline{X}}$ by $\underline{\underline{F}}$, estimates of the desired dependency relationships can now be based on a set of thoroughly orthogonal independent variables,

$$\underline{Y} = \underline{\underline{F}}\ \underline{\beta}^* + \underline{\epsilon}^* \ .$$

Taking advantage of the factor structure's internal orthogonality and, through the central limit theorem its approximate normality, each artificial variate's statistical significance can be tested with much greater confidence than econometric data ordinarily permits.

In some cases individual factors may be directly identified with meaningful economic phenomena through the subsets of variables that dominate their specifications. Should this be the case each component or factor may be interpreted and used as a variable in its own right, whose properties closely correspond to those required by the standard regression model. In such a case the transformation permits a reformulation of the model that can be tested effectively on existing data.

In general, however, the econometrician is not so fortunate; each factor turns out to be simply an artificial, linear combination of the original variables,

$$\underline{\underline{F}} = \underline{\underline{X}}\ \underline{\underline{A}}^t \ ,$$

that is completely devoid of economic content. In order to give meaning to structural coefficients, therefore, it is necessary to

return from factor to variable space by transforming estimators $\underline{A}$ and $\underline{\beta}^*$ into estimates

$$\underline{\beta}^{**} = \underline{\underline{A}}^t \underline{\beta}^*$$

of the structural parameters

$$\underline{Y} = \underline{\underline{X}} \underline{\beta} + \underline{u}$$

originally sought.

In the special (component analysis) case in which all $m = n$ factors are obtained, and retained in the regression equation. "[Nothing has been lost] by the transformation except the time spent on the arithmetical labor of finding it." † By the same token, however, nothing has been gained, for the Gaus-Markoff theorem insures that coefficient estimates $\underline{\beta}^{**}$ are identical to the estimates $\underline{\hat{\beta}}$ that would be obtained by the direct application of least squares to the original, highly unstable, set of variables. Moreover, all $m = n$ factors will be found significant and retained only in those instances in which the independent variable set, in fact, is not seriously multicollinear.

In general, therefore, Kendall's procedure derives n parameter estimates,

$$\underline{\beta}^{**} = \underline{\underline{A}}^t \underline{\beta}^* \ ,$$

from an m dimensional independent variable set,

$$\begin{aligned} \underline{Y} &= \underline{\underline{F}} \underline{\beta}^* + \underline{\epsilon}^* \\ &= (\underline{\underline{X}} - \underline{\underline{U}}) \underline{\underline{A}}^t \underline{\beta}^* + \underline{\epsilon}^* \ , \end{aligned}$$

† Kendall, op, cit., p. 70.

whose total information content is both lower and less well-defined than for the original set of variables. The rank of $\underline{\underline{X}}-\underline{\underline{U}}$, clearly, is never greater, and usually is smaller, than the rank of $\underline{\underline{X}}$. Multicollinearity, therefore, is intensified rather than alleviated by the series of transformations. Indeed, by discarding the residual -- or perhaps the "unique" -- portion of an independent variable's variation, one is seriously in danger of throwing out the baby rather than the bath -- i.e., the independent rather than the redundant dimensions of information.

Kendall's approach is not without attractions. Should factors permit identification and use as variables in their own right, the transformation provides a somewhat defensible solution to the multicollinearity problem. The discrepancy between apparent and significant dimensions (in model and data, respectively) is eliminated by a meaningful reduction in the number of the model's parameters. Even where factors cannot be used directly, their derivation provides insight into the pattern of interdependence that undermines the structural stability of estimates based on the original set of variables.

The shortcoming of this approach lies in its prescriptions for handling those situations in which the data do not suggest a reformulation that reduces the model's information requirements -- i.e., where components cannot be interpreted directly as economic variables. In such circumstances, solution of the multicollinearity problem requires the application of additional information, rather than the further reduction of existing information. Methods that retain a model's full complexity while reducing the information content of existing data aggrevate rather than alleviate the multicollinearity problem.

Rules of Thumb: A second and more pragmatic line of attack recognizes the need to live with poorly conditioned, non-experi-

mental data, and seeks to develop rules of thumb by which "acceptable" departures from orthogonality may be distinguished from "harmful" degrees of multicollinearity.

The term "harmful multicollinearity" is generally defined only symptomatically -- as the cause of wrong signs or other symptoms of nonsense regressions. Such a practice's inadequacy may be illustrated, perhaps, by the ease with which the same argument can be used to explain right signs and sensible regressions from the same basic set of data. An operational definition of harmful multicollinearity, however inadequate it may be, is clearly preferable to the methodological slight-of-hand that symptomatic definitions make possible.

The most simple, operational definition of unacceptable collinearity makes no pretense to theoretical validity. An admittedly arbitrary rule of thumb is established to constrain simple correlations between explanatory variables to less than, say, r = .8 or .9. The most obvious type of pairwise sample interdependence, of course, can be avoided in this fashion.

More elaborate and apparently sophisticated rules of thumb also exist. One, that has lingered in the background of econometrics for many years, has recently gained sufficient stature to be included in an elementary text. The rule holds, essentially, that "intercorrelation or multicollinearity is not necessarily a problem unless it is high relative to the over-all degree of multiple correlation..."* Or, more specifically, if

r_{ij} is the (simple) correlation between two independent variables, and

R_y is the multiple correlation between dependent and independent variables,

---

* L. R. Klein, An Introduction to Econometrics, Prentice-Hall, 1962; p. 101.

multicollinearity is said to be "harmful" if

$$r_{ij} \geq R_y \quad .$$

By this criterion the Cobb-Douglas production function is not seriously collinear, for multiple correlation $R_y = .98$ is comfortably greater than the simple correlation between (logarithms of) labor and capital, $r_{12} = .91$. Ironically this is the application chosen by the textbook to illustrate the rule's validity.*

Although its origin is unknown, the rule's intuitive appeal appears to rest on the geometrical concept of a triangle formed by the end points of three vectors (representing variables $\underline{Y}$, $\underline{X}_1$, and $\underline{X}_2$, respectively) in N dimensional observation space (reduced to three dimensions in Figure 1). $\underline{\hat{Y}} = \underline{X\hat{\beta}}$ is represented by the perpendicular (i.e., the least squares) reflection of $\underline{Y}$ onto the $\underline{X}_1$ $\underline{X}_2$ plane. Multiple correlation R_y is defined by the

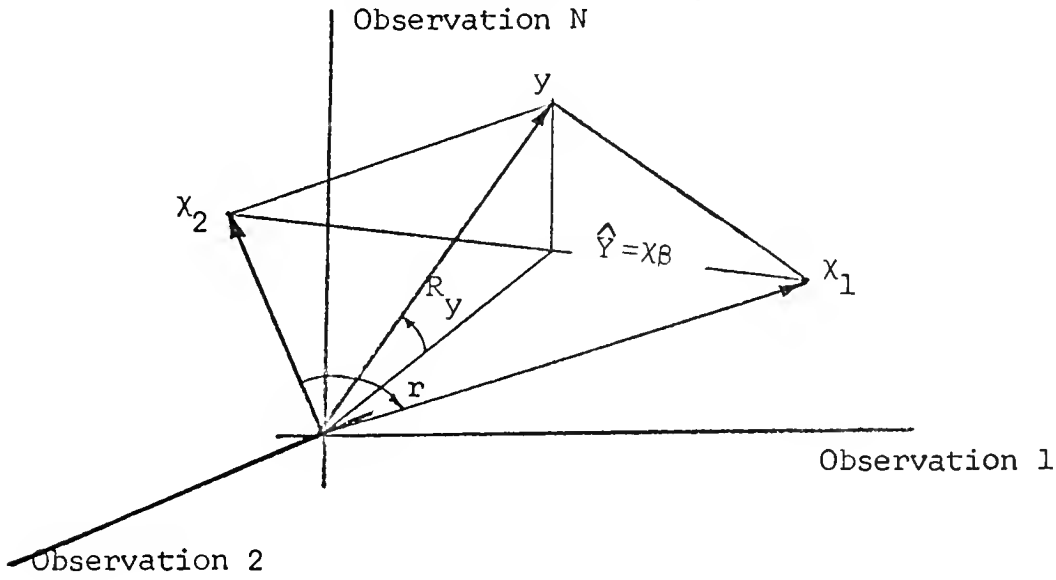


Figure 1

* Idem.

direction cosine between $\underline{Y}$ and $\hat{\underline{Y}}$, while simple correlation r_{12} is the direction cosine between $\underline{X}_1$ and $\underline{X}_2$. Should multiple correlation be greater than simple correlation, the triangle's base $\overline{X_1 X_2}$ is greater than its height $\overline{Y\,\hat{Y}}$, and the dependency relationship appears to be "stable."

Despite its intuitive appeal the concept may be attacked on at least two grounds. First, on extension to multiple dimensions it breaks down entirely. Complete multicollinearity -- i.e., perfect singularity -- within a set of explanatory variables is quite consistent with very small simple correlations between members of $\underline{\underline{X}}$. A set of dummy variables whose non-zero elements accidentally exhaust the sample space is an obvious, and an aggrevatingly common, example. Second, the Cobb-Douglas production function provides a convincing counter-example. If this set of data is not "harmfully collinear," the term has no meaning.

The rule's conceptual appeal may be rescued from absurdities of the first type by extending the concept of simple correlation between independent variables to multiple correlation within the independent variable set. A variable, $\underline{X}_i$, then, would be said to be "harmfully multicollinear" only if its multiple correlation with other members of the independent variable set were greater than the dependent variable's multiple correlation with the entire set.

The Cobb-Douglas counter example remains, however, to indicate that multicollinearity is basically an interdependency, not a dependency condition. Should $(\underline{\underline{X}}^t\underline{\underline{X}})$ be singular -- or virtually so -- tight sample dependence between $\underline{Y}$ and $\underline{\underline{X}}$ cannot assure the structural integrity of least squares parameter estimates.

Computer Programming

The development of large scale, high speed digital computers has had a well-recognized, virtually revolutionary impact

on econometric applications. By bringing new persons into contact with the field the computer also is having a perceptible, if less dramatic, impact on econometric methodology. The phenomenon is not new. Technical specialists have called attention to matters of theoretical interest in the past -- Professor Viner's famous draftsman, Mr. Wong, is a notable example.* More recently, the computer programmer's approach to singularity in regression analysis has begun to shape the econometrician's view of the problem as well.

Specifically, the numerical estimation of parameters for a standard regression equation requires the inversion of a matrix of variance-covariance or correlation coefficients for the independent variable set. Estimates of both slope coefficients,

$$\underline{\hat{\beta}} = (\underline{\underline{X}}^t \underline{\underline{X}})^{-1} \underline{\underline{X}}^t \underline{Y} \ ,$$

and variances,

$$\text{var}(\underline{\hat{\beta}}) = \sigma_u^2 \ (\underline{\underline{X}}^t \underline{\underline{X}})^{-1} \ ,$$

require the operation. Should the independent variable set $\underline{\underline{X}}$ be perfectly multicollinear, $(\underline{\underline{X}}^t \underline{\underline{X}})$, of course, is singular, and a determinate solution does not exist.

The programmer, accordingly, is required to build checks for non-singularity into standard regression routines. The test most commonly used relies on the property that the determinant of a singular matrix is zero. Defining a small, positive test value, $\epsilon > 0$, a solution is attempted only if the determinant

$$|\underline{\underline{X}}^t \underline{\underline{X}}| > \epsilon \ ;$$

otherwise, computations are halted and a premature exit is called.

* J. Viner, "Cost Curves and Supply Curves," Zeitschrift fur Nationalokonomie, III, 1931. Reprinted in A. E. A. Readings in Price Theory, Irwin, 1952.

Checks for singularity may be kept internal to a computer program, well out of the user's sight. Recently, however, the determinant has tended to join $\underline{\beta}$ coefficients, t-ratios, F-tests and other summary statistics as routine elements of printed output. Remembering that the determinant, $|\underline{\underline{X}}^t\underline{\underline{X}}|$, is based on a normalized, correlation matrix, its position on the scale

$$0 \leq |\underline{\underline{X}}^t\underline{\underline{X}}| \leq 1$$

yields at least heuristic insight into the degree of interdependence within the independent variable set. As $\underline{\underline{X}}$ approaches singularity, of course, $|\underline{\underline{X}}^t\underline{\underline{X}}|$ approaches zero. Conversely $|\underline{\underline{X}}^t\underline{\underline{X}}|$ close to one implies a nearly orthogonal independent variable set. Unfortunately, the gradient between extremes is not well defined. As an ordinal measure of the relative orthogonality of similar sets of independent variables, however, the statistic has attracted a certain amount of well-deserved attention and use.

A single, overall measure of the degree of interdependence within an independent variable set, although useful in its own right, provides little information on which corrective action can be based. Near singularity may result from strong, sample pairwise correlation between independent variables, or from a more subtle and complex linkage between several members of the set. The problem's cure, of course, depends on the nature of the interaction. The determinant per se, unfortunately, gives no information about that interaction.

In at least one case an attempt has been made to localize multicollinearity by building directly into a multiple regression program an index of each explanatory variable's dependence on other members of the independent variable set.* Recalling

* A. E. Beaton and R. R. Glauber, Statistical Laboratory Ultimate Regression Package, Harvard Statistical Laboratory, 1962.

again that $(\underline{\underline{X}}^t\underline{\underline{X}})$ is the matrix of simple correlation coefficients for $\underline{X}$, and in addition defining

r^{ij} as the i, j^{th} element of $(\underline{\underline{X}}^t\underline{\underline{X}})^{-1}$, and

$(\underline{\underline{X}}^t\underline{\underline{X}})_{ij}$ as the matrix of cofactors of the i, j^{th} element of $(\underline{\underline{X}}^t\underline{\underline{X}})$,

we have,

$$r^{ij} = \frac{(-1)^{i+j} \, |(\underline{\underline{X}}^t\underline{\underline{X}})_{ij}|}{|\underline{\underline{X}}^t\underline{\underline{X}}|} \quad .$$

Diagonal elements of the inverse, r^{ii}, accordingly may be represented as the ratio of determinants of two, positive semi-definite correlation matrices; the numerator containing each member of the independent variable set except $\underline{X}_i$, and the denominator containing the entire set $\underline{X}$.

Should $\underline{X}_i$ be orthogonal to other members of $\underline{X}$, $(\underline{\underline{X}}^t\underline{\underline{X}})_{ii}$ differs from $(\underline{\underline{X}}^t\underline{\underline{X}})$ only by the deletion of a row and column containing one on the diagonal and zeros off-diagonal. Thus,

$$|(\underline{\underline{X}}^t\underline{\underline{X}})_{ii}| = |\underline{\underline{X}}^t\underline{\underline{X}}| \ ,$$

and the diagonal element

$$r^{ii} = \frac{|(\underline{\underline{X}}^t\underline{\underline{X}})_{ii}|}{|\underline{\underline{X}}^t\underline{\underline{X}}|} = 1.$$

Should $\underline{X}_i$, on the other hand, be perfectly dependent on the other members of $\underline{X}$, the denominator $|\underline{\underline{X}}^t\underline{\underline{X}}|$ vanishes while the numerator $|(\underline{\underline{X}}^t\underline{\underline{X}})_{ii}|$, as it does not contain $\underline{X}_i$, is not affected. The

diagonal element r^{ii} -- and thereby the $\text{var}(\hat{\beta}_i) = \sigma_u^2 \; r^{ii}$ -- explodes, pinpointing not only the existence, but also the location of singularity within an independent variable set.

As originally conceived, diagonal elements of the inverse correlation matrix were checked internally by computer programs only to identify completely singular independent variables. More recently they have joined other statistics as standard elements of regression output.* Even though the spectrum $1 \leq r^{ii} \leq \infty$ is little explored, diagonal elements, by their size, give heuristic insight into the relative severity, as well as the location, of redundancies within an independent variable set.

Armed with such basic (albeit crude) diagnostics, the investigator may begin to deal with the multicollinearity problem. First, of course, the determinant $|\underline{\underline{X}}^t\underline{\underline{X}}|$ alerts him to its existence. Next, diagonal elements r^{ii} give sufficient insight into the problem's location -- and therefore into its cause -- to suggest the selective additions of information that are required for stable, least squares, parameter estimates.

---

* Idem.

THE PROBLEM REVISITED

Many persons, clearly, have examined one or more aspects of the multicollinearity problem. Each, however, has focused on one facet to the exclusion of others. Few have attempted to synthesize, or even to distinguish between either multicollinearity's nature and effects, or its diagnosis and cure. Klein, for example, defines the problem in terms that include both nature and effects; while Kendall attempts to produce a solution without concern for the problem's nature, effects, or diagnosis.

Those who do concern themselves with a definition of multicollinearity tend to think of the problem in terms of a discrete condition that either exists or does not exist, rather than as a continuous phenomenon whose severity may be measured.

A good deal of confusion -- and some inconsistency -- emerges from this picture. Cohesion requires, first of all, a clear distinction between multicollinearity's nature and effects, and, second, a definition in terms of the former on which diagnosis, and subsequent correction can be based.

DEFINITION

Econometric problems are ordinarily defined in terms of statistically significant discrepancies between the properties of hypothesized and sample variates. Non-normality, heteroscedasticity and autocorrelation, for example, are defined in terms of differences between the behavior of hypothesized and observed residuals. Such definitions lead directly to the development of test statistics on which detection, and an evaluation of the problem's nature and severity, can be based. Once an investigator is alerted to a problem's existence and character, of course, corrective action ordinarily constitutes a separate -- and often quite straight-forward -- step.

Such a definition would seem to be both possible and desirable for multicollinearity.

Let us define the multicollinearity problem, therefore, in terms of departures from parental orthogonality in an independent variable set. Such a definition has at least two advantages.

First, it distinguishes clearly between the problem's essential nature -- which consists of a lack of independence, or the presence of interdependence, in an independent variable set, $\underline{\underline{X}}$ -- and the symptoms or effects -- on the dependency relationship between $\underline{Y}$ and $\underline{\underline{X}}$ -- that it produces.

Second, parental orthogonality lends itself easily to formulation as a statistical hypothesis and, as such, leads directly to the development of test statistics, adjusted for numbers of variables and observations in $\underline{\underline{X}}$, against which the problem's severity can be calibrated. Developed in sufficient detail, such statistics may provide a great deal of insight into the location and pattern, as well as the severity, of interdependence that undermines the experimental quality of a given set of data.

DIAGNOSIS

Once a definition is in hand, multicollinearity ceases to be so inscruitable. Add a set of distributional properties and hypotheses of parental orthogonality can be developed and tested in a variety of ways, at several levels of detail. Statistics whose distributions are known (and tabulated) under appropriate assumptions, of course, must be obtained. Their values for a particular sample provide probabilistic measures of the extent of correspondence -- or non-correspondence -- between hypothe-

sized and sample characteristics; in this case, between hypothesized and sample orthogonality.

To derive test statistics with known distributions, specific assumptions are required about the nature of the population that generates sample values of $\underline{\underline{X}}$. Because existing distribution theory is based almost entirely on assumptions that $\underline{\underline{X}}$ is multivariate normal, it is convenient to retain the assumption here as well. Common versions of least squares regression models, and tests of significance based thereon, also are based on multivariate normality. Questions of dependence and interdependence in regression analysis, therefore, may be examined within the same statistical framework.

Should the assumption prove unnecessarity severe, its probabilistic implications can be relaxed informally. For formal purposes, however, multivariate normality's strength and convenience is essential, and underlies everything that follows.

General

The heuristic relationship between orthogonality and the determinant of a matrix of sample first order correlation coefficients

$$0 \leq |X^{t}X| \leq 1$$

has been discussed under computer programming approaches to singularity, above. Should it be possible to attach distributional properties under an assumption of parental orthogonality to the determinant $|\underline{\underline{X}}^{t}\underline{\underline{X}}|$, or to a convenient transformation of $|\underline{\underline{X}}^{t}\underline{\underline{X}}|$, the resulting statistic could provide a useful first measure of the presence and severity of multicollinearity within an independent variable set.

Presuming $\underline{\underline{X}}$ be multivariate normal, such properties are close at hand. As shown by Wishart, sample variances and covariances are jointly distributed according to the frequency function that now bears his name.* Working from the Wishart distribution, Wilkes, in an analytical tour de force, is able to derive the moments and distribution (in open form) of the determinant of sample covariance matrices.** Employing the additional assumption of parental orthogonality, he then obtains the moments and distribution of determinants for sample correlation matrices $|\underline{\underline{X}}^t\underline{\underline{X}}|$ as well. Specifically the k^{th} moment of $|\underline{\underline{X}}^t\underline{\underline{X}}|$ is shown to be

$$(2)\quad M_k\left(|\underline{\underline{X}}^t\underline{\underline{X}}|\right) = \frac{\left[\Gamma\left(\frac{N-1}{2}\right)\right]^{n-1} \prod_{i=2}^{n} \Gamma\left(\frac{N-i}{2} + k\right)}{\left[\Gamma\left(\frac{N-1}{2} + k\right)\right]^{n-1} \prod_{i=2}^{n} \Gamma\left(\frac{N-i}{2}\right)} \quad ,$$

where as before, N is sample size and n, the number of variables.***

In theory, one ought to be able to derive the frequency function for $|\underline{\underline{X}}^t\underline{\underline{X}}|$ from (2), and in open form it is indeed possible. For $n > 2$, however, explicit solutions for the distribution of $|\underline{\underline{X}}^t\underline{\underline{X}}|$ have not been obtained.

Bartlett, however, by comparing the lower moments of (2) with those of the Chi Square distribution, obtains a transformation of $|\underline{\underline{X}}^t\underline{\underline{X}}|$,

$$(3)\quad \dot{\chi}^2_{|\underline{\underline{X}}^t\underline{\underline{X}}|}(\nu) = -\left[N-1 - \frac{1}{6}(2n + 5)\right] \log |\underline{\underline{X}}^t\underline{\underline{X}}| \quad ,$$

---

* Wishart, J. "The Generalize Produce Moment Distribution in Samples from a Multivariate Normal Population," Biometrika, 20A 1928.

** Wilkes, S. "Certain Generalizations in the Analysis of Variance," Biometrika, 24, 1932; p. 477.

***Wilkes, S. op. cit., p. 492.

that is distributed approximately as Chi Square with $\nu = \frac{1}{2}n\ (n-1)$ degrees of freedom.

In this light the determinant of intercorrelations within an independent variable set takes on new meaning. No longer is interpretation limited to extremes of the range

$$0 \leq |\underline{\chi}^t\underline{\chi}| \leq 1 .$$

By transforming $|\underline{\chi}^t\underline{\chi}|$ into an approximate Chi Square statistic, a meaningful scale is provided against which departures from hypothesized orthogonality, and hence the gradient between singularity and orthogonoality, can be calibrated. Should one accept the multivariate normality assumption, of course, probability levels provide a cardinal measure of the extent to which $\underline{\chi}$ is interdependent. Even without such a scale, transformation to a variable whose distribution is known, even approximately -- by standardizing for sample size and number of variables -- offers a generalized, ordinal measure of the extent to which quite different sets of independent variables are undermined by multicollinearity.

Specific

Determining that a set of explanatory variables departs substantially from internal orthogonality is the first, but only the logical first step in an analysis of multicollinearity as defined here. If information is to be applied efficiently to alleviate the problem, localization measures are required to accurately specify the variables most severely undermined by interdependence.

To find the basis for one such measure we return to notions developed both by computer programmers and by econome-

tricians. As indicated earlier, both use diagonal elements of the inverse correlation matrix, r^{ii}, in some form, as measures of the extent to which particular explanatory variables are affected by multicollinearity.

Intuition suggests that our definition of hypothesized parental orthogonality be tested through this statistic.

Elements of the necessary statistical theory are developed by Wilkes, who obtains the distribution of numerous determinental ratios of variables from a multivariate normal distribution.* Specifically, for the matrix $(\underline{\underline{X}}^t\underline{\underline{X}})$, defining h principle minors

$$(4) \qquad |\underline{\underline{X}}^t\underline{\underline{X}}|_i \quad \text{for } i = 1 \ldots h,$$

such that no two contain the same diagonal element, r_{ii}, but that each r_{ii} enters one principle minor, Wilkes considers the variable

$$Z = \frac{|\underline{\underline{X}}^t\underline{\underline{X}}|}{\prod_{i=1}^{h} |\underline{\underline{X}}^t\underline{\underline{X}}|_i} \quad .$$

For any arbitrary set of h principle minors (4), he then obtains both the moments and distribution of Z. For the special case of interest here, (employing the notation of p.29 above), let us form principal minors such that

$h = 2$, $|\underline{\underline{X}}^t\underline{\underline{X}}|_1 = |r_{ii}| = 1$, and $|\underline{\underline{X}}^t\underline{\underline{X}}|_2 = |(\underline{\underline{X}}^t\underline{\underline{X}})_{ii}|$, then

$$Z_* = \frac{|\underline{\underline{X}}^t\underline{\underline{X}}|}{1 \times |(X^tX)_{ii}|} = \frac{1}{r^{ii}} \quad .$$

---

* S. S. Wilkes, op. cit., esp. pp. 480-2, 491-2.

Defining $\nu_1 = N - n$ and $\nu_2 = n - 1$, it follows from Wilkes' more general expression that the frequency distribution for Z_* can be written as,

$$(5)\quad f(Z_*) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\,\Gamma(\frac{\nu_2}{2})}\, Z_*^{\frac{1}{2}(\nu_1-2)}\,(1-Z_*)^{\frac{1}{2}(\nu_2-2)} \quad .$$

Now consider a change of variables to

$$(6)\quad \omega = (\frac{1}{Z_*} - 1)\,\frac{\nu_1}{\nu_2} = (r^{ii}-1)\,\frac{\nu_1}{\nu_2} \quad ,$$

and note that

$$(7)\quad Z_* = (\frac{\nu_2}{\nu_1}\,\omega + 1)^{-1} \quad ,$$

$$(8)\quad |\frac{dZ_*}{d\omega}| = (\frac{\nu_2}{\nu_1}\,\omega + 1)^{-2}\,(\frac{\nu_2}{\nu_1}) \quad ,$$

where the vertical bars in (8) denote absolute value. Substituting (7) into (5) and multiplying by (8), we have

$$g(\omega) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\,\Gamma(\frac{\nu_2}{2})}\,(\frac{\nu_2}{\nu_1}\,\omega + 1)^{\frac{1}{2}(2-\nu_1)}\,(1-[\frac{\nu_2}{\nu_1}\,\omega + 1]^{-1})^{\frac{1}{2}(\nu_2-2)}\,[(\frac{\nu_2}{\nu_1}\,\omega + 1)^{-2}(\frac{\nu_2}{\nu_1})]$$

$$= \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\,\Gamma(\frac{\nu_2}{2})}\,(\frac{\nu_2}{\nu_1}\,\omega + 1)^{-\frac{1}{2}(\nu_1+\nu_2)}\,\omega^{\frac{1}{2}(\nu_2-2)}\,(\frac{\nu_2}{\nu_1})^{\frac{1}{2}\nu_2} \quad ,$$

which can be recognized as the F-distribution with ν_1 and ν_2 degrees of freedom.*

The transformation

$$(9) \quad \omega = (r^{ii} - 1)\left(\frac{N-n}{n-1}\right) \quad ,$$

then, can be seen to be distributed as F with N-n and n-1 degrees of freedom. Defining $R^2_{\underline{\chi}_i}$ as the squared multiple correlation between $\underline{\chi}_i$ and the other members of $\underline{\chi}$, this result can be most easily understood by recalling that

$$r^{ii} = \frac{1}{1-R^2_{\underline{\chi}_i}} \quad .$$

Therefore, $(r^{ii}-1)$ equals $\dfrac{R^2_{\underline{\chi}_i}}{1-R^2_{\underline{\chi}_i}}$, and ω (as defined in (6) and (9) above), except for a term involving degrees of freedom, is the ratio of explained to unexplained variance; it is not surprising then, to see ω distributed as F.

As regards the distribution of (9), the same considerations discussed in the preceding section are relevant. If $\underline{\chi}$ is jointly normal, (9) is distributed exactly as F, and its magnitude therefore provides a cardinal measure of the extent to which individual variables in an independent variable set are affected by multicollinearity. If normality cannot be assumed (9) still provides an ordinal measure, adjusted for degrees of freedom, of $\underline{\chi}_i$'s dependence on other variables in $\underline{\chi}$.

Having established which variables in $\underline{\chi}$ are substantially

---

* F. Graybill, op. cit., p. 31.

multicollinear, it generally proves useful to determine in greater detail the pattern of interdependence between affected members of the independent variable set. An example, perhaps, will illustrate the information's importance. Suppose (9) is large only for $\underline{X}_1$, $\underline{X}_2$, $\underline{X}_3$, and $\underline{X}_4$, indicating these variables to be significantly multicollinear, but only with each other, the remaining variables in $\underline{\underline{X}}$ being essentially uncorrelated both with each other and with $\underline{X}_1$, ..., $\underline{X}_4$. Suppose further that all four variables, $\underline{X}_1$, ..., $\underline{X}_4$, are substantially intercorrelated with each of the others. If well-determined estimates are desired for this subset of variables, additional information must be obtained on at least three of the four.

Alternatively, suppose that $\underline{X}_1$ and $\underline{X}_2$ are highly correlated, $\underline{X}_3$ and $\underline{X}_4$ also are highly correlated, but all other intercorrelations among the four, and with other members of $\underline{\underline{X}}$, are small. In this case, additional information must be obtained only for two variables -- $\underline{X}_1$ or $\underline{X}_2$, and $\underline{X}_3$ or $\underline{X}_4$. Clearly, then the efficient solution of multicollinearity problems requires detailed information about the pattern as well as the existence, severity, and location of intercorrelations within a subset of interdependent variables.

To gain insight into the pattern of interdependence in $\underline{\underline{X}}$, a straight forward transformation of off-diagonal elements of the inverse correlation matrix $(\underline{\underline{X}}^t\underline{\underline{X}})^{-1}$ is both effective and convenient. Its development may be summarized briefly, as follows.

Consider a partition of the independent variable set

$$\underline{\underline{X}} = \begin{pmatrix} \underline{\underline{X}}^{(1)} \\ \underline{\underline{X}}^{(2)} \end{pmatrix}$$

such that variables $\underline{X}_i$ and $\underline{X}_j$ constitute $\underline{\underline{X}}^{(1)}$, and the remaining

n-2 variables $\underline{\underline{X}}^{(2)}$. The corresponding matrix of zero order correlation coefficients, then, is partitioned such that

$$(\underline{\underline{X}}^t\underline{\underline{X}}) = \begin{pmatrix} \underline{\underline{R}}_{11} & \underline{\underline{R}}_{12} \\ \underline{\underline{R}}_{21} & \underline{\underline{R}}_{22} \end{pmatrix}$$

where $\underline{\underline{R}}_{11}$, containing variables $\underline{X}_i$ and $\underline{X}_j$, is of dimension 2 x 2 and R_{22} is (n-2) x (n-2). Elements of the inverse correlation matrix $\underline{\underline{r}}^{ij}$ corresponding to $\underline{\underline{X}}^{(1)}$, then, can be expressed without loss of generality as*

$$\underline{\underline{r}}^{ij} = (\underline{\underline{R}}_{11} - \underline{\underline{R}}_{12}\ \underline{\underline{R}}_{22}^{-1}\ \underline{\underline{R}}_{21})^{-1} .$$

Before inversion the single off-diagonal element of

$$(\underline{\underline{R}}_{11} - \underline{\underline{R}}_{12}\ \underline{\underline{R}}_{22}^{-1}\ \underline{\underline{R}}_{21})$$

may be recognized as the partial covariance of $\underline{X}_i$ and $\underline{X}_j$, holding constant $\underline{\underline{X}}^{(2)}$, the other members of the independent variable set. On normalizing -- i.e., dividing by square roots of corresponding diagonal elements -- in the usual fashion, partial correlation coefficients between $\underline{X}_i$ and $\underline{X}_j$ can be obtained.**

For the special case considered here, where $\underline{\underline{X}}^{(1)}$ contains only 2 variables and $\underline{\underline{R}}_{11}$, accordingly is 2 x 2, it can also be shown that corresponding normalized off-diagonal elements of $(\underline{\underline{R}}_{11} - \underline{\underline{R}}_{12}\ \underline{\underline{R}}_{22}^{-1}\ \underline{\underline{R}}_{21})$ and its inverse $(\underline{\underline{R}}_{11} - \underline{\underline{R}}_{12}\ \underline{\underline{R}}_{22}^{-1}\ \underline{\underline{R}}_{21})^{-1}$ differ

---

* G. Hadley, Linear Algebra, Addison Wesley, 1961; pp. 107, 108.

** T. W. Anderson, An Introduction to Multivariate Statistical Analysis, Wiley, 1958.

from one another only by sign. It follows, therefore, that, by a change of sign, normalized off-diagonal elements of the inverse correlation matrix $(\underline{\underline{X}}^t\underline{\underline{X}})^{-1}$ yield partial correlations among members of the independent variable set. That is, defining $r_{ij\cdot}$ as the coefficient of partial correlation between $\underline{X}_i$ and $\underline{X}_j$, other members of $\underline{\underline{X}}$ held constant, and r^{ij} as elements of $(\underline{\underline{X}}^t\underline{\underline{X}})^{-1}$, above, it follows that

$$r_{ij\cdot} = \frac{-r^{ij}}{\sqrt{r^{ii}}\sqrt{r^{jj}}}.$$

Distributional properties under a hypothesis of parental orthogonality, of course, are required to tie-up the bundle. Carrying forward the assumption of multivariate normality, such properties are close at hand. In a manner exactly analogous to the simple (zero order) correlation coefficient, the statistic

$$t_{ij\cdot}(\nu) = \frac{r_{ij\cdot}\sqrt{N-n}}{\sqrt{1-r^2_{ij\cdot}}}$$

may be shown to be distributed as Student's t with $\nu = N-n$ degrees of freedom.*

An exact, cardinal interpretation, of interdependence between $\underline{X}_i$ and $\underline{X}_j$ as members of $\underline{\underline{X}}$, of course, requires exact satisfaction of multivariate normal distributional properties. As with the determinant and diagonal elements of $(\underline{\underline{X}}^t\underline{\underline{X}})^{-1}$ that precede it, however, off-diagonal elements -- transformed to $r_{ij\cdot}$ or $t_{ij\cdot}$ -- provide useful ordinal measures of collinearity even in the absence of such rigid assumptions.

---

* Graybill, op. cit., pp. 215, 208.

Illustration

A three stage hierachy of increasingly detailed tests for the presence, location, and pattern of interdependence within an independent variable set $\underline{\underline{X}}$ has been proposed. In order, the stages are:

1. Test for the presence and severity of multicollinearity anywhere in $\underline{\underline{X}}$, based on the approximate distribution (3) of determinants of sample correlation matrices, $|\underline{\underline{X}}^t \underline{\underline{X}}|$, from an orthogonal parent population.
2. Test for the dependence of particular variables on other members of $\underline{\underline{X}}$ based on the exact distribution, under parental orthogonality, of diagonal elements of the inverse correlation matrix $(\underline{\underline{X}}^t \underline{\underline{X}})^{-1}$.
3. Examine the pattern of interdependence among $\underline{\underline{X}}$ through the distribution, under parental independence, of off-diagonal elements of the inverse correlation matrix, $(\underline{\underline{X}}^t \underline{\underline{X}})^{-1}$.

In many ways such an analysis, based entirely on statistics that are routinely generated during standard regression computations, may serve as a substitute for the formal,thorough (and time-consuming) factor analysis of an independence variable set. It provides the insight required to detect, and if present to identify, multicollinearity in $\underline{\underline{X}}$. Accordingly, it may serve as a starting point from which the additional information required for stable, least squares estimation can be sought. An illustration, perhaps, will help to clarify the procedure's mechanics and purpose; both of which are quite straight-forward.

In a series of statistical cost analyses for the U.S. Navy, an attempt has been made to measure the effect on maintenance cost of such factors as ship age, size, intensity of usage (measured by fuel consumption), time between successive overhauls, and such discrete, qualitative characteristics as propulsion mode (steam,

diesel, nuclear), complexity (radar picket, guided missile, etc.), and conversion under a recent (Fleet Rehabilitation and Modernization, FRAM) program. Equations have been specified and estimated on various samples from the Atlantic Fleet destroyer force that relate logarithms of repair costs to logarithms of age, displacement, overhaul cycle and fuel consumption, and to discrete (0, 1) dummy variables for diesel propulsion, radar picket, and FRAM conversion.*

Stability under changes in specification, direction of mimization, and sample coverage have been examined heuristically by comparing regression coefficients, determinants of correlation matrices $|\underline{\underline{X}}^t \underline{\underline{X}}|$, and diagonal elements of $(\underline{\underline{X}}^t \underline{\underline{X}})^{-1}$, from different equations. The sensitivity of certain parameters under such changes (e.g., fuel consumption) and the stability of others (e.g., age, overhaul cycle) have been noted in the past.

By performing an explicit analysis for interdependence in $\underline{\underline{X}}$, such information could have been obtained more quickly, directly, and in greater detail. Consider, for example. the seven variable equation summarized in Table 3. Multiple correlation and associated F-statistics, with t-ratios for the relationship between dependent and independent variables, shows dependence between $\underline{Y}$ and $\underline{\underline{X}}$ to be substantial.

---

* D. E. Farrar and R. E. Apple, "Some Factors that Affect the Overhaul Cost of Ships," Naval Research Logistics Quarterly, 10, 4, 1963; and ..... "Economic Considerations in Establishing an Overhaul Cycle for Ships," Naval Engineers Journal, 77, 6, 1964.

Overhaul Cost Equation, U. S. Navy Destroyers*

Measures of Dependence

$R_y^2 = .80$ $F_y(7,88) = 56$

$$\text{Log } Y = 4.81 + .34 \text{ Log } X_1 + .40 \text{ Log } X_2 - .79 \text{ Log } X_3 + .05 \text{ Log } X_4 - .30 X_5 + .11 X_6 - .16$$

$t(88) =$ 11.4, 5.4, 8.7, 0.5, 3.5, 2.5, 2.

Measures of Interdependence

$\dot{\chi}^2 |\underline{\chi}^t \underline{\chi}| \ (21) = 261$

$F_{\underline{\chi}_i}(6,89) =$ 6.3, 19.9, 6.3, 47.5, 47.1, 12.7, 4.

Pattern of Interdependence

Partial $r_{ij.}$ below diagonal - Multiple $R^2_{\underline{\chi}_i}$ on diagonal - Partial $t_{ij.}$ above diagonal.

	Age	Size		Cycle	Fuel	Diesel	Radar	FRAM
Age	.30	1.27	$t_{ij.}$	2.01	-3.57	1.27	4.72	3.38
Size	.13	.57		2.60	3.44	-2.03	.78	-1.27
Cycle	.21	.27	R_1^2	.30	.82	1.15	-3.51	.59
Fuel	-.35	.34		.09	.76	-8.77	3.13	4.06
Diesel	-.27	-.21	$Y_{ij.}$	.12	-.68	.76	5.51	2.68
Radar	.45	.08		-.35	.31	.50	.46	-2.55
FRAM	.34	-.13		.06	.40	.27	-.26	.24

Sample size, N = 96

Y = Overhaul cost (thousands of dollars)

X_1 = Age (years)

X_2 = Size (displacement, thousands of tons)

X_3 = Overhaul cycle (years)

X_4 = Fuel consumption (standardized)

X_5 = 1 if diesel propulsion
= 0 if not

X_6 = 1 if radar picket
= 0 if not

X_7 = 1 if FRAM
= 0 if not

* D. E. Farrar and R. E. Apple, op. cit.

Measures of interdependence within $\underline{\underline{X}}$, beginning with the approximate Chi Square transformation for the matrix of correlation coefficients over the entire set,

$$\dot{\chi}^2 | \underline{\underline{X}}^t \underline{\underline{X}} | \ (21) = 261$$

quickly alert one, however, to the existence of substantial muli-collinearity in $\underline{\underline{X}}$, as well.

Multiple correlations and associated F-statistics within $\underline{\underline{X}}$ -- to measure each explanatory variable's dependence on other members of the set -- shows $\underline{X}_1$, $\underline{X}_3$, $\underline{X}_7$, (age, cycle and FRAM) to be quite stable; $\underline{X}_2$, $\underline{X}_6$ (size and radar picket) to be moderately affected by multicollinearity; and $\underline{X}_4$, $\underline{X}_5$ (fuel consumption and diesel propulsion) to be extremely multicollinear.

Off-diagonal partial correlations and associated t-ratios show a complex linkage involving fuel consumption, diesel propulsion, and radar picket to lie at the heart of the problem.

The next step is up to the model builder. Should his purpose be to provide forecasts or to suggest policy changes that require reliable information about structural relationships between repair cost and either age or overhaul frequency, the job already is done. Dependence between $\underline{Y}$ and $\underline{X}_1$, $\underline{X}_3$ is strong, and interdependence between these (explanatory) variables and other members of $\underline{\underline{X}}$ is weak. The experimental quality of this portion of our data is high and estimates, accordingly, are likely to be stable.

Should one's purpose be to make forecasts or policy proposals that require accurate knowledge of the link between repair cost and fuel consumption, propulsion mode, or radar picket, on the other hand, corrective action to obtain more substantial information is required. In this particular case sample stratification may help to overcome the problem. In other instances more strenuous efforts -- such as additional primary data collection, extraneous

parameter estimates from secondary data sources, or the direct application of subjective information -- may be necessary.

In any case, efficient corrective action requires selectivity, and selectivity requires information about the nature of the problem to be handled. The procedure outlined here provides such information. It produces detailed diagnostics that can support the selective acquisition of information required for effective treatment of the multicollinearity problem.

SUMMARY

A point of view as well as a collection of techniques is advocated here. The techniques -- in this case a series of diagnostics -- can be formulated and illustrated explicitly. The spirit in which they are developed, however, is more difficult to convey. Given a point of view, techniques that support it may be replaced quite easily; the inverse is seldom true. An effort will be made, therefore, to summarize our approach to multicollinearity and to contrast it with alternative views of the problem.

- Multicollinearity as defined here is a statistical, rather than a mathematical condition. As such one thinks, and speaks, in terms of the problem's severity rather than of its existence or non-existence.

- As viewed here, multicollinearity is a property of the independent variable set alone. No account whatever is taken of the extent, or even the existence, of dependence between $\underline{Y}$ and $\underline{\underline{X}}$. It is true, of course, that the effect on estimation and specification of interdependence in $\underline{\underline{X}}$ -- reflected by variances of estimated regression coefficients -- also depends partly on the strength of dependence between $\underline{Y}$ and $\underline{\underline{X}}$. In order to treat the problem, however, it is important to distinguish between nature and effects, and to develop diagnostics based on the former. In our view an independent variable set $\underline{\underline{X}}$ is not less multicollinear if related to one dependent variable than if related to another; even though its effects may be more serious in one case than the other.

- Of multicollinearity's effects on the structural integrity of estimated econometric models -- estimation instability, and structural misspecification -- the latter, in our view, is

the more serious. Sensitivity of parameter estimates to changes in specification, sample coverage, etc., is reflected at least partially in standard deviations of estimated regression coefficients. No indication at all exists, however, of the bias imparted to coefficient estimates by incorrectly omitting a relevant, yet multicollinear, variable from an independent variable set.

Historical approaches to multicollinearity are almost unanimous in presuming the problem's solution to lie in deciding which variables to keep and which to drop from an independent variable set. Thought that the gap between a model's information requirements and data's information content can be reduced by increasing available information, as well as by reducing model complexity, is seldom considered.*

A major aim of the present approach, on the other hand, is to provide sufficiently detailed insight into the location and pattern of interdependence among a set of independent variables that strategic additions of information become not only a theoretically possibility, but also a practically feasible solution for the multicollinearity problem.

- Selectivity, however, is emphasized. This is not a counsel of perfection. The purpose of regression analysis is to estimate the structure of a dependent variable $\underline{Y}$'s dependence on a pre-selected set of independent variables $\underline{X}$, not to select an orthogonal independent variable set.**

---

* H. Theil, op cit, p. 217; and J. Johnston, op cit, p. 207 are notable exceptions.

** Indeed, should a completely orthogonal set of economic variables appear in the literature one would suspect it to be either too small to explain properly a moderately complex dependent variable, or to have been chosen with internal orthogonality rather than relevance to the dependent variable in mind.

Structural integrity over an entire set, admittedly, requires both complete specification and internal orthogonality. One cannot obtain reliable estimates for an entire n dimensional structure, or distinguish between competing n dimensional hypotheses, with fewer than n significant dimensions of independent variation. Yet all variables are seldom equally important. Only one -- or at most two or three -- strategically important variables are ordinarily present in a regression equation. With complete specification and detailed insight into the location and pattern of interdependence in $\underline{X}$, structural instability within the critical subset can be evaluated and, if necessary, corrected. Multicollinearity among non-critical variables can be tolerated. Should critical variables also be affected additional information to provide coefficient estimates either for strategic variables directly, or for those members of the set on which they are principally dependent -- is required. Detailed diagnostics for the pattern of interdependence that undermines the experimental quality of $\underline{X}$ permits such information to be developed and applied both frugally and effectively.

- Insight into the pattern of interdependence that affects an independent variable set can be provided in many ways. The entire field of factor analysis, for example, is designed to handle such problems. Advantages of the measures proposed here are twofold. The first is pragmatic; while factor analysis involves extensive separate computations, the present set of measures relies entirely on transformations of statistics, such as the determinant $|\underline{X}^t \underline{X}|$ and elements of the inverse correlation matrix, $(\underline{X}^t\underline{X})^{-1}$, that are generated routinely during standard regression computations. The second is symmetry; questions of dependence and interdependence in regression analysis are handled in the same conceptual and statistical framework. Variables that are internal to a set $\underline{X}$ for one purpose are viewed as external to it for another. In

this vein, tests of interdependence are approached as successive tests of each independent variable's dependence on other members of the set. The conceptual and computational apparatus of regression analysis, accordingly, is used to provide a quick and simple, yet serviceable, substitute for the factor analysis of an independent variable set.

- It would be plesant to conclude on a note of triumph that the problem has been solved and that no further "revisits" are necessary. Such a feeling, clearly, would be misleading. Diagnosis, although a necessary first step, does not insure cure. No miraculous "instant orthogonalization" can be offered.

We do, however, close on a note of optimism. The diagnostics described here offer the econometrician a place to begin. In combination with a spirit of selectivity in obtaining and applying additional information, multicollinearity may return from the realm of impossible to that of difficult, but tractable, econometric problems.